HW6_STAT5000

Xingyu Chen

November 03, 2021

Total Score: 30/32

1 Theoretical Questions

1. The weekly production of a corn farm can be modeled with a normal random variable that has a mean of 8 tons and a variance of 4 tons.

(a) What is the probability that, in a given week, production is greater than 11 tons? Q1: 2/2

Answer:

$Z = \frac{X-\mu}{\sigma} = \frac{11-8}{2} = 1.5$

$P(X > 11) = 1 - \Phi(1.5) = 1 - 0.933 = 0.067$

(b) What is the probability that, in a given week, production is between 6 and 7.5 tons? Q2: 2/2

Answer:

$Z_1 = \frac{X-\mu}{\sigma} = \frac{6-8}{2} = -1$

$Z_2 = \frac{X-\mu}{\sigma} = \frac{7.5-8}{2} = -0.25$

$P(6 < X < 7.5) = \Phi(-0.25) - \Phi(-1) = 0.401 - 0.159 = 0.242$

(c) How many tons represents the 35th percentile in weekly production? Q3: 2/2

Answer:

$P(X < P_{35}) = P(Z < \frac{P_{35}-\mu}{\sigma} = 0.35)$

$\frac{P_{35}-\mu}{\sigma} = -0.385$

$P_{35} = \mu - 0.385 * \sigma = 8 - 0.385 * 2 = 7.23$ tons

2.

(a) Suppose that 10 people in a sample of 85 are smokers. Calculate the 95% confidence interval for the true proportion of smokers in the population. Q4: 2/2

Answer:

$n = 85$

$p = 10/85 = 0.1263$

$p \pm Z\sqrt{pq/n} = 0.12 \pm 1.96\sqrt{0.12 * 0.88/85} = 0.12 \pm 0.0691 = (0.0509, 0.1891)$

(b) In 1881 Michelson and Newcomb measured the time light took to travel a distance of 7400 meters. From a a study of their experimental setup and a descriptive study of their 64 measurements and experimental setup, we conclude that the data can be assumed to be iid. These measurements yield the following sample quantities in μs (sec $\times 10^{-6}$), also known as microseconds:

$$\bar{x} = 27.75, \quad s = 5.08$$

Construct an approximate 95% confidence interval for the time light takes to travel 7400 meters. Q5: 2/2

Answer:
$\bar{x} \pm Z\sigma/\sqrt{n} \Rightarrow 27.75 \pm 1.96 * 5.08/\sqrt{64} = 27.75 \pm 1.2446 = (26.5054, 28.9946)$

3. Let$Z = \frac{X-\mu}{\sigma}$,where$X \sim \mathcal{N}(\mu, \sigma^2)$. Prove that E(Z) = 0 and Var(Z) = 1. Q6: 2/2

Answer:
$Z = \frac{X-\mu}{\sigma}$

$E(Z) = E(\frac{X-\mu}{\sigma})$
$E(Z) = (1/\sigma) * E(X - \mu)$
$E(Z) = (1/\sigma) * (E(X) - E(\mu))$
$E(Z) = (1/\sigma) * 0$
$E(Z) = 0$

$Var(Z) = Var(\frac{X-\mu}{\sigma})$
$Var(Z) = (1/\sigma)^2 * Var(X - \mu)$
$Var(Z) = (1/Var(X)) * (Var(X) - Var(\mu))$
$Var(Z) = (1/Var(X)) * (Var(X) - 0))$
$Var(Z) = Var(X)/Var(X)$
$Var(Z) = 1$

4. Let $\widehat{\theta}$ be an estimator of the parameter θ (e.g., we might think of $\widehat{\theta} = \bar{X}$ and $\theta = \mu$, where μ is a population mean). We say that $\widehat{\theta}$ is unbiased if $E(\widehat{\theta}) = \theta$.

(a) Let $X_1, ..., X_n$ be an iid sample from a population with mean μ and variance σ^2. Show that $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ is an unbiased estimator of . This answers the question of why we divide by $n-1$ in s^2 ! Q7: 2/2

Answer:

$$E\left(S^2\right) = E\left[\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}\right] = \frac{1}{n-1}E\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

$$= \frac{1}{n-1}E\sum_{i=1}^{n}\left(X_i^2 + \bar{X}^2 - 2\bar{X}X_i\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}E\left(X_i^2\right) - nE\left(\bar{X}^2\right)\right]$$

$E\left(X_i^2\right) = \mu^2 + \sigma^2 \quad E\left(\bar{X}^2\right) = \mu^2 + \sigma^2/n$

$$E\left(S^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2+\sigma^2\right) - n\left(\mu^2+\sigma^2/n\right)\right]$$
$$= \frac{1}{n-1}\left(n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2\right)$$
$$= \sigma^2$$

(b) Assume that $E(\sqrt{X}) \leq \sqrt{E(X)}$. Show that s is a biased estimator of σ. Q8: 2/2

Answer: $S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

$E(S) = E\left(\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}\right) = E\left(\sqrt{S^2}\right) \leq \sqrt{E\left(S^2\right)} = \sqrt{\sigma^2} = \sigma$

So S is biased estimate of σ

2 Computational Questions

Instructions for "computational" questions: Your work should be neatly done and in- clude all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do not put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade.

1. Load hubble.txt into R. A description of the variables can be obtained from page 73 of https://cran.r-project.org/web/packages/gamair/gamair.pdf.

(a) Calculate the 85% confidence interval for the mean of a galaxy's distance from Earth in Mega parsecs in R by doing the computation explicitly. Q9: 2/2

Answer:

```
library(gamair)
library(tidyverse)
hubble_df <- get(data(list= 'hubble'))
mean_df <- mean(hubble_df[[3]])
sd_df <- sd(hubble_df[[3]])
mean_df - 1.06 * sd_df / sqrt(24)
```

```
## [1] 10.79646
```

```
mean_df + 1.06 * sd_df / sqrt(24)
```

```
## [1] 13.31271
```

(b) Can you find a built in R function that does this computation automatically? Q10: 2/2

Answer:

```
t.test(hubble_df[[3]], conf.level = 0.85)
```

```
##
```

```
##  One Sample t-test
##
## data:  hubble_df[[3]]
## t = 10.156, df = 23, p-value = 5.701e-10
## alternative hypothesis: true mean is not equal to 0
## 85 percent confidence interval:
##  10.28695 13.82222
## sample estimates:
## mean of x
##  12.05458
```

(c) Interpret the confidence interval. Q11: 2/2

Answer: There is a 85% chance that the confidence interval I calculated (10.28695, 13.82222) contains the true population mean.

2. Simulate m = 1, 000 90% confidence intervals for a population proportion p (using the confidence interval formula that we derived in class). Use this simulation to justify the interpretation of this confidence interval. Q12: 0/2

Answer:

```
daf <- sample(1000)
mean_df <- mean(daf)
sd_df <- sd(daf)
mean_df - 1.64 * sd_df / sqrt(1000)
```

```
## [1] 485.5214
```

```
mean_df + 1.64 * sd_df / sqrt(1000)
```

```
## [1] 515.4786
```

```
t.test(daf, conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  daf
## t = 54.8, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  485.4632 515.5368
## sample estimates:
## mean of x
##     500.5
```

3. We might be interested in computing confidence intervals for parameters other than a mean, μ, proportion, p, etc. For many of these parameters, standard statistical theory will not help. In this problem, we will compute a 95% confidence interval for the rate

parameter of an exponential distribution.
A theoretical model suggests that X, the time to breakdown of an insulating fluid between electrodes at a particular voltage, has an exponential distribution: $f(x;\lambda) = \lambda e^{-\lambda x}$. A random sample of n = 10 breakdown times (minutes) is given here:

$$41.53, 18.73, 2.99, 30.34, 12.33, 117.52, 73.02, 223.63, 4, 26.78.$$

(a) Construct a matrix of B = 10, 000 rows, where each row is a sample of size 10 (sampled with replacement) from the above 10 numbers. (HINT: use the sample function in R.) Q13: 2/2

Answer:

```
value <- c(41.53, 18.73, 2.99, 30.34, 12.33, 117.52, 73.02, 223.63, 4, 26.78)
result <- matrix(sample(value, size = 100000, replace = TRUE), nrow = 10000, ncol = 10)
head(result)
```

```
##        [,1]  [,2]  [,3]  [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
## [1,]  30.34 18.73  4.00 73.02   4.00 223.63   4.00 223.63   4.00 117.52
## [2,]   2.99 73.02 26.78 73.02  30.34   4.00  18.73 223.63  41.53   2.99
## [3,] 223.63  2.99  2.99 41.53  12.33 117.52  18.73 117.52 117.52  26.78
## [4,]   4.00 30.34  2.99 41.53 117.52  26.78 117.52  73.02 117.52 117.52
## [5,] 223.63 73.02 73.02 26.78 117.52  41.53 117.52  12.33  41.53  41.53
## [6,]  73.02  2.99  2.99 26.78   4.00  30.34  18.73  30.34   4.00  41.53
```

(b) From each of the B samples, compute a reasonable estimate of λ (HINT: How is λ related to the mean of an exponential?). Call this estimator $\widehat{\lambda}$. Q14: 2/2

Answer:

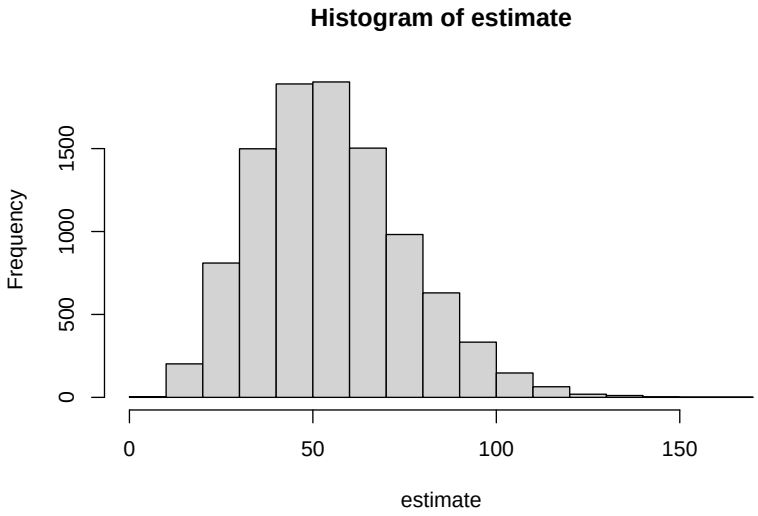
```
estimate <- rowMeans(result)
head(estimate)
```

```
## [1] 70.287 49.703 68.154 64.874 76.841 23.472
```

(c) You now have a sample from the distribution of the estimator $\widehat{\lambda}$. Construct a histogram and comment on the distribution. Q15: 2/2

Answer:

```
hist(estimate)
```



It is normal distribution.

(d) Use the quantile function in R to find the 2.5 percentile and the 97.5 percentile. This is a bootstrap confidence interval for λ. Q16: 2/2

Answer:

```
quantile(estimate, probs = c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 20.90097 99.87570
```